Grenze International Journal of Engineering and Technology, January Issue

Sentiment Analysis Empowered Drug Recommendations

Katta Naga Priyanka[1], Dudyala Akash[2], Parthi Jayadeep[3] and Shilpa Bagade[4]
[1-3]Electronics and Communication Engineering, Gokaraju Rangaraju Institute of Engineering and Technology Hyderabad, India
Email: priyakat42@gmail.com, ak2911dud@gmail.com, jayadeep2020@gmail.com
[4]Assistant Professor, Electronics and Communication Engineering, Gokaraju Rangaraju Institute of Engineering and Technology, Hyderabad, India
Email: shilpa.me1437@gmail.com

***Abstract*—The process of gathering, detecting, and monitoring information regarding side effects, adverse reactions, and precautions of pharmaceutical products poses significant challenges. With the emergence of user forums, online reviews have become a crucial source of product information. The outbreak of flu, which has led to fatalities, has caused disruption in the medical community, prompting individuals to self-medicate due to limited access to medical consultation, exacerbating their health issues. Recent advancements in machine learning and automation have shown promise across various applications. This study aims to propose a strategy for prescribing medications that can alleviate the workload of specialists. A medication recommendation system is developed in this project, utilizing techniques such as Bag of Words (Bow), TF-IDF, Word2Vec, and manual feature analysis, along with patient ratings to predict sentiment. Various classification algorithms are employed to suggest the most suitable medication for a specific ailment. The sentiment analysis of drug reviews is studied using machine learning classifiers, including Logistic Regression, Perceptron, Multinomial Naive Bayes, Decision Tree, Random Forest, LGBM, and CatBoost applied to Word2Vec, as well as the manual features approach. Additionally, classifiers utilizing Bow, TF-IDF, and Bow using SVC and stochastic gradient descent are utilized. The evaluation metrics include f1-score accuracy, precision, recall, and AUC score. The data-set used in the drug recommendation system based on sentiment analysis of drug reviews comprises drug names, drug sentiment, and drug ratings.**

***Index Terms*— Bow, TF-IDF, Word2Vec, Logistic Regression.**

I. INTRODUCTION

The number of instances of coronary sclerosis is skyrocketing, and the medical crisis is most acute in rural regions, where there are fewer doctors than in cities. Between the ages of six and twelve, doctors must complete their schooling. As a result, rapidly increasing the number of doctors is impossible. In this trying time, the telemedicine framework must be as effective as possible [1].

Today, clinical mistakes are common. Medication mistakes harm 100,000 people in the US and over 200,000 people in China annually. Experts frequently prescribe incorrect medications (more than 40% of the time) because they have little knowledge . [2]. For patients who need a doctor with an extensive understanding of

Grenze ID: 01.GIJET.11.1.425

patients, antimicrobial, and microscopic organisms, choosing the correct prescription is crucial [3]. Every day, fresh medications and diagnostic tools for medical professionals are released, along with new studies. It becomes more challenging for doctors to select a patient's therapy or medication based on indications and clinical history.
Reviews of articles have become more significant as a result of the Internet's explosive growth and the rise of the web-based business sector. People across the world are now accustomed to reading reviews and browsing websites before making purchases. The majority of earlier research has focused on evaluating expectations and recommendations for the e-commerce sector; health care and pharmaceuticals have only sporadically been explored. Because of their health worries, more people are now looking for online diagnoses.
Nearly 60% of Americans searched for health-related subjects online in 2013, while roughly 35% looked for medical information, per a 2013 Pew American Research Centre poll [4]. The pharmaceutical recommender's structure is crucial for assisting physicians and patients in comprehending the medicine for certain medical issues.
A popular approach is the recommendation framework, in which the user suggests products based on his or her benefits and requirements[5]. Utilizing consumer surveys, these frameworks analyse1QA the responses and provide suggestions depending on the precise requirements of the respondents.. he drug recommendation system employs psychological analysis and functional engineering to give drugs under circumstances depending on patient assessments. Sentiment analysis is a technique of recognizing and extracting emotional information from texts such under certain circumstances. On the contrary, by adding existing features through a feature engineering process, the performance of the model is increased. Five parts make up this study project: [6].The importance of this research is succinctly outlined in the opening section. An overview of earlier research in this area is given in the associated section. In the methodology section, the approaches employed in this study are thoroughly explained. In the result section, the outcomes of the applied model are assessed using a variety of parameters, and then debates and conclusions are drawn from the study.

II. Literature Survey

The use of advanced techniques like deep learning and machine learning has revolutionized recommendation systems in various industries such as travel, e-commerce, and dining, leading to improved user experiences. However, the application of these methods in drug recommendation systems has not been widely explored due to the complexity involved in analyzing medication reviews. These reviews often contain specialized clinical terms and pharmaceutical language, making sentiment analysis a challenging task. The specific vocabulary used in healthcare presents unique difficulties for sentiment analysis when compared to other fields. Despite the potential benefits, there is a noticeable lack of research in this area, highlighting the need for advancements in AI to address this gap.[7]
The study presents GalenOWL, a semantic-empowered online framework, to help specialists discover details on the medications. The paper depicts a framework that suggests drugs for a patient based on the patient's infection, sensitivities, and drug [8] interactions. For empowering GalenOWL, clinical data, and terminology are first converted to ontological terms utilizing worldwide standards, such as ICD-10 and UNII, and then correctly combined with the clinical information.
Leilei Sun [9] conducted research on extensive treatment records to identify the most suitable treatment plans for patients. The approach involved utilizing a proficient semantic clustering algorithm to gauge the resemblances among treatment records. Additionally, the author developed a framework to evaluate the effectiveness of the recommended treatments. This framework has the capability to recommend optimal treatment plans for new patients based on their demographics and medical conditions. Testing of this framework utilized Electronic Medical Records (EMR) collected from multiple clinics. The findings indicate that this framework enhances the rate of successful treatment outcomes.
The study conducted multilingual sentiment analysis by employing Naive Bayes and Recurrent Neural Network (RNN). The Google Translator API was utilized to translate multilingual tweets into English. The findings revealed that RNN achieved superior performance with an accuracy of 95.34%, outperforming Naive Bayes, which achieved an accuracy of 77.21%.
The study [11] is because the recommended drug should depend upon the patient's capacity. For example, if the patient's immunity is low, at that point, reliable medicines ought to be recommended. Proposed a risk level classification method to identify the patient's immunity. For example, more than 60 risk factors, hypertension, liquor addiction, and so forth have been adopted, which decide the patient's capacity to shield himself from infection. A web-based prototype system was also created, which uses a decision support system that helps doctors select first-line drugs.

Xiaohong Jiang et al. [12] Three separate algorithms, namely the decision tree algorithm, support vector machine (SVM), and back propagation neural network, were evaluated using treatment data. Among them, SVM was selected for the medication recommendation module due to its consistently high performance across three distinct metrics: model accuracy, model efficiency, and model robustness (R versatility). Additionally, a error checking mechanism was proposed to ensure analysis, accuracy, and management quality.
Mohammad Mehedi Hassan et al. [13] developed a cloud-assisted drug proposal (CADRE). As per patients' side effects, CADRE can suggest drugs with top-N related prescriptions. This proposed framework was initially founded on collaborative filtering techniques in which the medications are initially bunched into clusters as indicated by the functional description data. However, after considering its weaknesses computationally costly, cold start, and information sparsity, the model is shifted to a cloud-helped approach using tensor decomposition to advance the quality of experience of medication suggestion.
Considering the significance of hashtags in sentiment analysis, Jiugang Li et al. [14] A framework for recommending hashtags was developed, employing the skip-gram model and constitutional neural networks (CNN) to generate semantic sentence vectors. These vectors were then utilized to classify hashtags using LSTM recurrent neural networks (RNN). Findings show that this model outperforms traditional approaches such as SVM and standard RNN. This investigation highlights that conventional AI methods like SVM and collaborative filtering techniques tend to overlook semantic features, which significantly impacts prediction accuracy.

III. METHODOLOGY

The data-set used in this research is the Drug Review Data-set (ibm.com) taken from the UCI ML repository. This data set contains six attributes, age (numerical), sex (text) of a patient, BP (text) of a patient, cholesterol (text) of a patient, Na_to_K(numerical), and a drug to be used according to the patient's condition.

A. Decision Tree

For supervised learning tasks like classification and regression, the decision tree technique is a popular option. It functions by building a structure like a tree, where each node represents an option that was made in response to the objective numbers and characteristics in the dataset. The decision-making process is simple to comprehend and analyze thanks to this hierarchical paradigm. Decision trees are useful for a wide range of applications because they are capable of handling both categorical and numerical information well.

B. Splitting Criterion

Using criteria such as the Gini Index, Entropy, and Information Gain, the decision tree algorithm determines the optimal feature for splitting the data. The Gini Index assesses the impurity of a node by examining the likelihood of misidentifying a randomly chosen element within it. In contrast, Entropy quantifies the disorder or uncertainty of class distribution within a node to measure its impurity. Information Gain evaluates the reduction in entropy achieved by splitting a node based on a specific attribute, indicating the improvement in purity. These criteria are crucial for determining the feature to use when partitioning the data, thereby aiding in the creation of an effective decision tree model.

C. Recursive Binary Splitting

A root node which includes the whole datasets is where the decision tree algorithm starts. Then, depending on predefined attributes and threshold values, it splits the node into child nodes recursively. A feature and a splitting threshold are marked by each internal node, and a class label or projected value is indicated by every single leaf node. The question of whether a child node's feature values fall below or above a specified threshold affects how data instances are distributed to them. Recursively dividing data is carried out until a stopping condition is satisfied, like reaching the maximum depth or a certain minimum quantity of sample data per leaf node. By going through this procedure, the decision tree algorithm creates a hierarchical structure that effectively facilitates tasks like regression or classification by identifying the fundamental patterns in the data.

D. Pruning

Pruning is a crucial technique employed to mitigate the complexity of decision trees and counteract over fitting. This process entails the removal of branches or nodes from the tree structure that fail to substantially enhance the model's performance on unseen data. Various pruning methods exist, among which two prominent approaches are cost-complexity pruning and reduced error pruning. Cost-complexity pruning involves iteratively pruning nodes from the tree while considering a complexity parameter, which balances the simplicity of the tree with its

accuracy. On the other hand, reduced error pruning involves systematically evaluating the impact of pruning nodes on the overall error rate of the model, selectively removing nodes that lead to minimal increases in error. By effectively applying pruning techniques, decision trees can achieve better generalization performance and robustness, thereby enhancing their utility in real-world applications.

E. Prediction

To generate predictions for new instances, the decision tree algorithm traverses the tree structure starting from the root node and moving down to a leaf node. This traversal is guided by the feature values of the instance being evaluated. At each internal node, the algorithm compares the feature value of the instance with the splitting threshold associated with that node and proceeds to the appropriate child node based on this comparison. This process continues recursively until a leaf node is reached. The class label assigned to this leaf node represents the predicted class for classification tasks, while for regression tasks, the value associated with the leaf node serves as the predicted outcome. By following this traversal mechanism, the decision tree algorithm effectively predicts the class or value for new instances based on the learned patterns encoded in the tree structure.

F. Mathematical Equations Of The Algorithm

GINI Impurity

Gini impurity: The Gini impurity measures the impurity or the probability of misidentifying a randomly selected element from a set of data points.
Gini impurity equation:
Gini(p) = 1 - Σ (pi)^2
Where: Gini(p) is the Gini impurity for a given node.
Σ represents the summation.
pi is the probability of class i.

Entropy

Entropy: Entropy measures the impurity or disorder in a set of data points.
Entropy equation:
Entropy(D) = - Σ (pi) * log2(pi)
Where: Entropy(D) is the entropy of the dateset D.
Σ represents the summation. pi is the probability of class I

Implementation

Model Construction

Importing modules and downloaded packages

1. Numpy
2. Pandas
3. Seaborn
4. Matplotlib.pyplot
5. DecisionTreeClassifier from sklearn.tree

Importing Packages And Reading Datasets

The number of rows, column names, non-null count, and data types are among the details that are printed in this line about the data frame. These Create the feature matrix X in lines by choosing columns from my Data Frame. The values property is used to pick the columns "Age,"
"Sex," "BP," "Cholesterol," and "Na to K" and turn them into a NumPy array. The first five rows of X are printed in the second line.

Data Preprocessing

These By choosing the 'Drug' column from my data Data Frame, generate the target vector Y in lines. The first five values of Y are displayed on the second line. Utilizing the Label Encoder () class from sklearn's preprocessing, encode the category features in the feature matrix X into numerical values in lines. The features "Sex," "BP," and "Cholesterol" are each given a label encoder, totaling three. The matching columns in X are transformed once the label encoders have been fitted to the category values. The modified values are applied to X once more. The altered X's first five rows are printed in the final line.

Setting Up The Decision Tree

1.**Import Libraries**: Bring in necessary libraries, particularly `DecisionTreeClassifier` from scikit-learn.

2. **Prepare Data**: Divide your data into features (X) and the target variable (y).
3.**Create Model**: Set up a `DecisionTreeClassifier` instance.
4. **Train Model**: Fit the model to your training data using fit()`.
5.**Make Predictions**: Utilize the trained model to predict outcomes for new data with `predict()`.
6.**Evaluate Model (Optional)**: Gauge the model's performance using metrics such as accuracy.
7. **Visualize Tree (Optional)**: Visualize the decision tree structure to gain insights.

Predicition and Evaluation

Once your machine learning model is trained on historical data, you apply it to new, unseen data for predictions. In Python, following model fitting on the training data, you employ the `predict()` method to generate predictions on new data sets.

In Python, post making predictions on a test datasets, you assess your ML model's performance using metrics like accuracy, precision, recall, and F1-score for classification tasks, and metrics like mean squared error (MSE), mean absolute error (MAE), and R-squared for regression tasks. Libraries like scikit-learn provide functions for calculating these metrics, making evaluation straightforward.

System Architecture

To make sure a system or application fits the needs of its users, system designers and developers can use a basic architecture diagram (UML) to show the high-level structure of their system or application. This term may also be used to describe the patterns present in the design. It serves as an example of a technique used to specify restrictions, connections, and boundaries between components while abstracting the general layout of a software system. The software system's physical deployment and evolution plan are detailed in great detail. The designers and developers who use this diagram stand to gain a lot from it.

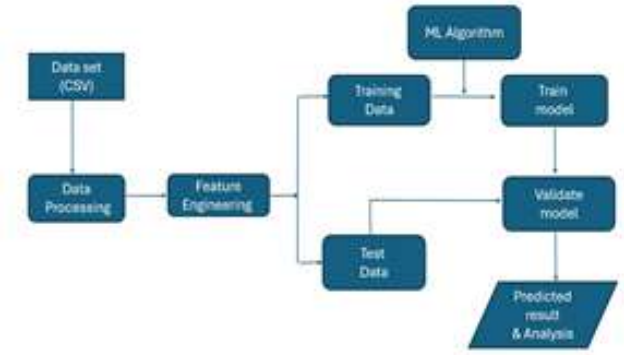

Fig.1 System Architecture

Use-Case Diagram

Demonstrates relationships between actors (users or external systems) and use cases (system features) to describe a system's functioning from the point of view of the user.

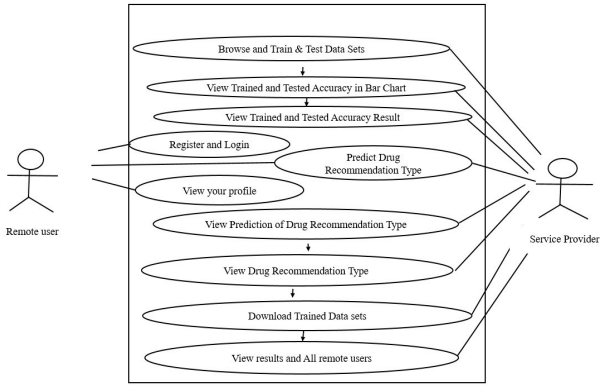

Fig.2 Use-Case Diagram

IV. RESULT

The graphic representation of the decision tree model's structure is shown in the figure. Each node in the tree indicates a judgment or a condition based on one of the attributes. The nodes are connected by branches that depict several outcomes or routes, depending on the criteria. The decision tree has a root node at the top and branches out to several internal nodes and leaf nodes. Leaf nodes reflect the ultimate anticipated class or outcome, whereas internal nodes include the decision rules or conditions. The node's color and shading reveal the distribution of classes or the dominant class for a certain node. Entropy is used by the decision tree to divide the data into sections and calculate the information gain at each node. The tree is arranged hierarchically, with the most specific nodes at the bottom and the maximum information gain at the top.

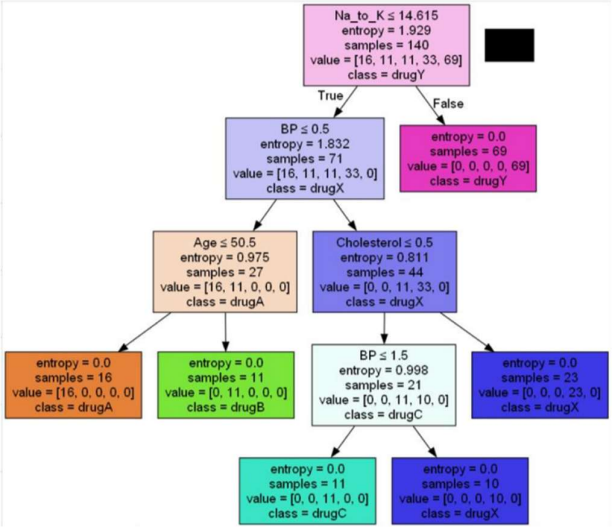

Fig.3 Decision Tree Algorithm

The graphic representation of the decision tree model's structure is shown in the figure. Each node in the tree indicates a judgment or a condition based on one of the attributes. The nodes are connected by branches that depict several outcomes or routes, depending on the criteria. The decision tree has a root node at the top and branches out to several internal nodes and leaf nodes. Leaf nodes reflect the ultimate anticipated class or outcome, whereas internal nodes include the decision rules or conditions. The node's color and shading reveal the distribution of classes or the dominant class for a certain node. Entropy is used by the decision tree to divide the data into sections and calculate the information gain at each node. The tree is arranged hierarchically, with the most specific nodes at the bottom and the maximum information gain at the top.

TABLE-I

Model	Accuracy	Precision	Recall	F-1 score
Logistic Regression	0.5667	05681	0.5666	0.5658
LinearSVC	0.5667	0.5680	0.5667	0.5660
Multinomia|NB	0.5670	05682	0.5668	0.5666
Randomforest Classifier	0.5670	0.5678	0.5667	0.5662
LGBM classifier	0.5666	0.5686	0.5668	0.5662
Catboost classfier	0.5668	0.5680	0.5666	0.5660
Decision Tree	0.9833	0.9840	0.9833	0.9831

The decision tree comes out as the model that performs the best overall among those mentioned. With a high accuracy of 0.9833, it successfully predicts the class labels for the majority of the dataset's cases. With a rating of 0.9840, the decision tree exhibits outstanding accuracy as well as a very low percentage of false positive predictions. Furthermore, the model has a high recall of 0.9833, indicating that it correctly detects positive cases while avoiding numerous false negatives. The decision tree thus gets a remarkable F1-score of 0.9831, which is a harmonic mean of accuracy and recall and denotes a well-balanced performance. Across all criteria, the performance of the other models, which include Logistic Regression, linear SVC, multinomial NB, random forest classifier, LGBM classifier, and catboost classifier, is comparatively comparable. These models have

accuracy values between 0.5666 and 0.5670, precision between 0.5678 and 0.5686, recall between 0.5666 and 0.5668, and F1-scores between 0.5658 and 0.5666. According to these results, these models perform about as well as the decision tree model in terms of accuracy, precision, recall, and F1-score.

V. CONCLUSION

By providing patients and healthcare professionals with tailored, evidence-based recommendations, pharmaceutical recommendation systems have the potential to significantly enhance healthcare outcomes. These systems combine extensive patient data sets, medical knowledge, and cutting-edge algorithms to generate tailored recommendations that take into account the patient's features, medical history, drug interactions, and suggested course of treatment. But for the successful deployment and usage of medicine recommendation systems, rigorous algorithm validation, careful consideration of data privacy and security, and good integration into clinical workflows are all required. These tools should only ever be used as decision support tools, leaving healthcare professionals' clinical judgement and domain knowledge to make the final decisions. Despite the fact that these tools can enhance decisionmaking and promote patient safety. Drug recommendation systems need to advance in order to be reliable, accurate, and practical in real healthcare settings. To achieve this, it is necessary for technology creators and healthcare professionals to work together and conduct constant research and reviews.

REFERENCE

[1] Wittich CM, Burkle CM, Lanier WL. Medication errors: an overview for clinicians. Mayo Clin Proc. 2014 Aug;89(8):1116-25.

[2] CHEN, M. R., & WANG, H. F. (2013). The reason and prevention of hospital medication errors. Practical Journal of Clinical Medicine, 4.

[3] Fox, Susannah & Duggan, Maeve. (2012). Health Online 2013. Pew Research Internet Project Report.

[4] Maeve Duggan, Fox, and Susannah."Health Online 2013."thefollowingURLhttp://pewinternet.org/Reports/2013/Health-online.aspx

[5] Bartlett, John G., et al. "Practice Guidelines for the Management of Community-Acquired Pneumonia in Adults." *Clinical Infectious Diseases*, vol. 31, no. 2, Oxford UP (OUP), Aug. 2000, pp. 347–82. *Crossref*, https://doi.org/10.1086/313954.

[6] Bartlett, John G., et al. "Practice Guidelines for the Management of Community-Acquired Pneumonia in Adults." *Clinical Infectious Diseases*, vol. 31, no. 2, Oxford UP (OUP), Aug. 2000, pp. 347–82. *Crossref*, https://doi.org/10.1086/313954.

[7] "Probabilistic Aspect Mining Approach for Interpretation and Evaluation of Drug Reviews." *IEEE Conference Publication | IEEE Xplore*, 1 Oct. 2016, doi.org/10.1109/SCOPES.2016.7955684.

[8] Doulaverakis, Charalampos, et al. "GalenOWL: Ontology-based Drug Recommendations Discovery." *Journal of Biomedical Semantics*, vol. 14, no. 1, 1 Jan. 2012, https://doi.org/10.1186/2041-1480-3-14.

[9] Sun, Leilei, et al. *Data-driven Automatic Treatment Regimen Development and Recommendation*. 13 Aug. 2016, https://doi.org/10.1145/2939672.2939866.

[10] "Sentiment Analysis of Multilingual Twitter Data Using Natural Language Processing." *IEEE Conference Publication | IEEE Xplore*, 1 Nov. 2018, doi.org/10.1109/CSNT.2018.8820254.

[11] Shimada, K., Takada, H., Mitsuyama, S., Ban, H., Matsuo, H., Otake, H., ... & Kaku, M. (2005). Drug-recommendation system for patients with infectious diseases. In AMIA Annual Symposium Proceedings (Vol. 2005, p. 1112). American Medical Informatics Association.

[12] "An Intelligent Medicine Recommender System Framework." *IEEE Conference Publication | IEEE Xplore*, 1 June 2016, doi.org/10.1109/ICIEA.2016.7603801.

[13] Zhang, Yin, et al. "CADRE: Cloud-Assisted Drug Recommendation Service for Online Pharmacies." *Mobile Networks and Applications*, vol. 348–355, no. 3, 10 Dec. 2014, https://doi.org/10.1007/s11036-014-0537-4.

[14] "Tweet Modeling With LSTM Recurrent Neural Networks for Hashtag Recommendation." *IEEE Conference Publication | IEEE Xplore*, 1 July 2016, doi.org/10.1109/IJCNN.2016.7727385.

[15] Zhang, Yin, et al. "Understanding Bag-of-words Model: A Statistical Framework." *International Journal of Machine Learning and Cybernetics*, vol. 43–52, no. 1–4, 28 Aug. 2010, https://doi.org/10.1007/s13042-010-0001-0.

[16] Chawla, Nitesh V., et al. "SMOTE: Synthetic Minority Over-sampling Technique." *Journal of Artificial Intelligence Research*, vol. 321–357, 1 June 2002, https://doi.org/10.1613/jair.953.

[17] *Bioinfo Publications - Peer Reviewed Academic Journals*. doi.org/10.9735/2229-3981.

[18] "ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning." *IEEE Conference Publication | IEEE Xplore*, 1 June 2008, doi.org/10.1109/IJCNN.2008.4633969.

[19] "SMOTETomek-Based Resampling for Personality Recognition." *IEEE Journals & Magazine | IEEE Xplore*, 2019, doi.org/10.1109/ACCESS.2019.2940061.